ATTY. DOCKET NO. 50P4487

UNITED STATES PATENT APPLICATION

FOR

RETRAINING AND UPDATING SPEECH MODELS FOR SPEECH RECOGNITION

INVENTORS:

Courtney Konopka
and
Lars Cristian Almstrand

CERTIFICATE OF EXPRESS MAILING

"Express Mail" Label: EL645987231 US
Date of Deposit: August 16, 2001
I hereby certify that the above-referenced application papers are being deposited with the United States Postal Service "Express Mail Post Office to Addressee" service under 37 C.F.R. §1.10 on the date indicated above and is addressed to the Commissioner for Patents, Washington, D.C. 20231.

Carrie Merzbacher

RETRAINING AND UPDATING SPEECH MODELS FOR SPEECH RECOGNITION

BACKGROUND OF THE INVENTION

This invention relates to a method and apparatus for updating speech models that are used for speech recognition systems and, more particularly, to a technique for the creation of new, or retraining of existing speech models, to be used to recognize speech from a class of users whose speech differs sufficiently from the modeled language of the system that speaker adaptation is not feasible. In this document, the term "speech models" refers collectively to those components of a speech recognition system that reflect the language which the system models, including for example, acoustic speech models, pronunciation entries, grammar models. In a preferred implementation of the present invention, speech models stored in a central repository, or database, are updated and then redistributed to participating client sites where speech recognition is carried out. In another embodiment, the repository may be co-located with a speech server, which in turn recognizes speech sent to it, as either processed (i.e. derived speech feature vectors), or unprocessed speech.

Speech recognition systems convert the utterance of a user into digital form and then process the digitized speech in accordance with known algorithms to recognize the words and/or phrases spoken by the user. For example, in co-pending application serial No. (450103-02964) a speech recognition system is disclosed wherein the digitized speech is processed to extract sequences of feature sets that describe corresponding speech passages. The speech passage is then recognized by matching the corresponding feature set sequence with the optimal model sequence.

The process of selecting the models to compare to the features derived from the utterance are constrained by prior rules which limit the sets of models and their corresponding configurations (patterns) that are used. It is these selected models and patterns that are used to recognize the user's speech. These rules include grammar rules and lexica by which the statistically most probable speech model is selected to identify the words spoken by the user. A conventional modeling

approach that has been used successfully is known as the hidden Markov model (HMM). While HMM modeling has proven successful, it is difficult to employ a universal set of speech models successfully. Speech characteristics vary between speaker groups, within speaker groups and even for single individuals (due to health, stress, time, etc.), to the extent that a preset speech model may need adaptation or retraining to best adapt itself to the utterances of a particular user.

Such correction and retraining is relatively simple when adapting speech models of a user whose speech matches the data used to train the speech recognition system, because speech from the user and the training group have certain common characteristics. Hence, relatively small modifications to a preset speech model to adapt from those common characteristics are readily achievable, but large deviations are not. Various accents, inflections, pathological speech or other speech features contained in the utterances of such an individual are sufficiently different from the preset speech models as to inhibit successful adaptation retraining of those models. For example, the acoustic subwords pronounced by users whose primary language is not the system target language are quite different from the target language acoustic subwords to which the speech models of typical speech recognition systems are trained. In general, subwords pronounced by "non-native" speakers typically exhibit a transitional hybrid between the primary language of those users and target language subwords. In another example, brain injury, or injury or malformation of the physical speech production mechanism, can significantly impair a speaker's ability to pronounce certain acoustic subwords in conformance with the speaking population at large. A significant subgroup if this speech-impaired population would require the speech models such a system would create.

Ideally, speech recognition systems should be trained with data that closely models the speech of the targeted speaker group, rather that adapted from a more general speech model. This is because it is a simpler, more efficient task, having a higher probability of success, to train uniquely designed speech models to recognize the utterances of such users, rather than correct and retrain preset system target-language models. However, the creation of uniquely designed speech models is time-consuming in and of itself and requires an large library of speech data and subsequently models that are particularly representative of, and adapted to several

different speaker classes. Such data-dependency poses a problem, because for HMM's to be speaker independent, each HMM must be trained with a broad representation of users from that class. But, if an HMM is trained with overly broad data, as would be the case for adapting to speech the two groups of speakers exemplified above, the system will tend to misclassify the features derived from that speech.

This problem can be overcome by training HMM's less broadly, and then adapting those HMM's to the utterances of a specific speaker. While this approach would reduce the error rate (i.e. the rate of misclassification) for some speakers, it is of limited utility for certain classes of speakers, such as users whose language is not a good match with the system target language.

Another approach is to train many narrow versions of HMM's, each for a particular class of users. These versions may be delineated in accordance with various factors, such as the primary language of the user, a particular geographic region of the user's dialect, the gender of the user, his or her age, and so on. When combined with speaker adaptation, that is, the process of adapting an HMM to best match the utterances of a particular speaker, this approach has the potential to produce speech models of the highest accuracy. However, since there are so many actual and potential classes of users, a very large database of training data, (and subsequently speech models) would be needed. In addition, since spoken language itself is a dynamic phenomenon, the system target language speech models (lexicon, acoustic models, grammar rules, etc.) and sound system change over time to reflect a dynamic speaking population. Consequently, the library of narrow HMM versions would have to be corrected and retrained continually in order to reflect those dynamic aspects of speaking population at large. This would suggest a repository to serve as a centralized point for the collection of speech data, training of new models and the redistribution of these improved models to participating users, i.e. a kind of "language mirror".

SUMMARY OF INVENTION

Therefore, it is an object of the present invention to provide a technique that is useful in recognizing speech from users whose spoken language differs from the primary language of the typical speech recognition system.

Another object of this invention is to provide a technique for correcting, retraining and updating speech models that are used in speech recognition systems to best recognize speech that differs from that of such speech recognition systems.

A further object of this invention is to provide a central repository of speech models that are created and/or retrained in accordance with the speech of a particular class of users, and then distributed to remote sites at which speech recognition is carried out.

Yet another object of this invention is to provide a system for efficiently and inexpensively adapting speech recognition models to diverse classes of users, thereby permitting automatic speech recognition of utterances from those users.

Various other objects, advantages and features of the present invention will become readily apparent from the ensuing detailed description, and the novel features will be particularly pointed out in the appended claims.

In accordance with this invention, a technique is provided for updating speech models for speech recognition by identifying, from a class of users, acoustic subword (e.g. phoneme) data for a predetermined set of utterances that differ from a set of stored speech models by at least a predetermined amount. The identified subword data for similar utterances from the class of users is collected and used to correct the set of stored speech models. As a result, the corrected speech models are a closer match to the utterances than were the set of stored speech models. The set of speech models are subsequently updated with the corrected speech models to provide improved speech recognition of utterances from the class of users. Further, a technique is provided for identifying preferred pronunciations for said class.

As another feature of this invention, speech models are built for recognizing speech of users of a particular class by registering users in accordance with predetermined criteria that characterize the speech of the class, sensing an utterance from a user, determining a best match of the utterance to a stored speech model and collecting utterances from users of the particular class that differ from the stored, best match speech model by at least a predetermined amount. The stored speech model then is retrained to reduce to less than the predetermined amount the

difference between the retrained speech model and the identified utterances from users of the particular class.

BRIEF DESCRIPTION OF THE DRAWINGS

The following detailed description, given by way of example, and not intended to limit the present invention solely thereto, will best be understood in conjunction with the accompanying drawings in which:

Fig. 1 is a block diagram of one example of a system which incorporates the present invention;

Figs. 2A-2B constitute a flow chart of the manner in which the system depicted in Fig. 1 operates; and

Figs. 3A-3B constitute a more detailed flow chart depicting the operation of the present invention.

DETAILED DESCRIPTION OF A PREFERRED EMBODIMENT

Referring now to Fig. 1,there is illustrated a block diagram of the speech recognition system in which the present invention finds ready application. The system includes user sites 10_1, 10_2, ... 10_n, each having a user input 12 and speech recognition apparatus 14; and a remote, central site having a central database 18 and a speech recognition processor module 20 that operates to correct and retrain speech models stored in central database 18. User input 12 preferably includes a suitable microphone or microphone array and supporting hardware of the type that is well known to those of ordinary skill in the art. One example of a suitable user input is described in aforementioned application (attorney docket 450100-02964).

User input 12 is coupled to speech recognition apparatus 14 which operates to recognize the utterances spoken by the user and supplied thereto by the user input. The speech recognition apparatus preferably includes a speech processor, such as may be implemented by a suitably programmed microprocessor, operable to extract from the digitized speech samples identifiable speech features. This feature data from the user are identified and compared to stored speech models such as phonetic models, implemented as a library of, for example, HMM's. The optimal match of the user's feature data to a stored model results in identifying the

corresponding utterance. As is known, one way to match an utterance to a stored model is based upon the Viterbi score of that model. The Viterbi score represents the relative level of confidence that the utterance is properly recognized. It will be appreciated that the higher the Viterbi score, relative to the other models in the system, the higher the level of confidence that the corresponding utterance has been properly recognized.

As is known, one way to determine whether an utterance is a member of the vocabulary modeled by the system is based on the rejection score of the utterance. In one embodiment, the rejection score is calculated as the frame-normalized ratio of the in-vocabulary (IV) Viterbi score over the out-of-vocabulary (OOV) Viterbi score. The IV score is calculated as the best Viterbi score calculated by the system for the utterance. The OOV score is calculated by applying the utterance to an ergotically connected network of all acoustic subword (ASW) models. As is known, this sort of network is called an ergotically connected model.

When the rejection score of an utterance is above a predetermined threshold, but is less than a score representing absolute confidence, that is, when the rejection score is in a "gray" area, the system and/or user will recognize when the attempted identification of the utterance is incorrect and will have an opportunity to correct the identification of that utterance. Over a period of time, corrections made by the system and/or user will result in an updating, or retraining of the stored speech models to which the user's utterances are compared. Such updating or retraining is known to those of ordinary skill in the art; one example of which is described in the text, "Fundamentals of Speech Recognition" by Lawrence Rabiner and Biing-Wang Juang. Also, the Baum-Welch algorithm is a standard technique for establishing and updating models used in speech recognition. However, such techniques are of limited utility if, for example, the primary language spoken by the user differs from the primary language to which the speech recognition system is programmed. For instance, if the primary language of the user is not the system target language, or if the dialect spoken by the user is quite different from the dialect recognized by the system, the updating and retraining of speech models (that is, the "learning" process) is not likely to succeed. The objective of the present invention is to provide a technique that is ancillary to the conventional speech recognition system to update

and retrain the typical HMM's or other acoustic speech models for subsequent successful speech recognition of utterances from such users.

In accordance with this invention, user input 12 is provided with, for example, a keyboard or other data input device, by which the user may enter predetermined criteria that characterize his speech as being spoken by users of a particular class. That is, the user input is operable to enter class information. Criteria that identify a user's class may include, but are not limited to, the primary language spoken by the user, the user's gender, the user's age, the number of years the user has spoken the system target language , user's height, user's weight, the age of the user when he first learned the system target language, and the like. It is clear that any criteria that characterizes the user may be specified. In addition, samples of calibrated utterances of the user are entered by way of user input 12. Such calibrated utterances are predetermined words, phrases and sentences which are compared to the stored speech models in speech recognition apparatus 14 for determining the rejection score of those calibrated utterances and for establishing the type and degree of correction needed to conform the stored speech models to those utterances. All of this information, referred to as class data, is used to establish and register the class of the user. For example, the class of the user may be determined to be French male, age 35, the system target language spoken for 15 years; or Japanese female, age 22, the system target language learned at the age of 14 and spoken for eight years; or an Australian male. This class data is transferred from the user's site via a suitable communications channel, such as an Internet connection, a telephonic or wireless media 16, to database 18 and the remote central site. Such data transfer could either occur in an interactive mode, whereby said data transfer would take place immediately, and whereby the user would wait for a response from the server, or occur in a batch mode, where said data transfer would occurs when data traffic on the communications channel is relatively light, such as during early morning hours.

Speech recognition apparatus 14 at the user's site also identifies those utterances that are not satisfactorily recognized, such as those utterances having low confidence levels. For example, acoustic subword data that differs from a best-matched speech model by at least a predetermined amount, that is, subword data whose rejection score is within a particular range, referred to above as the gray area,

is identified. The corresponding best-matched speech model and the correction data needed to correct that speech model to achieve a closer match to the user's utterance are accumulated for later transmission to database 18 if batch processing is implemented, or forwarded immediately, if interactive mode is implemented. In either mode, the data is transferred to database 18 whereat it is associated with the class registered by the user. Consequently, over time, database 18 collects utterances, speech models and correction data from several different users, and this information is stored in association with the individual classes that have been registered. Hence, identified subword data for similar utterances from respective classes together with the rejection scores of those utterances and the correction data for those subwords are collected. After a sufficient quantity, or so-called "critical mass", of subword and correction data has been collected at database 18, module 20 either creates, or retrains the speech models stored in the database, resulting in an updated set of speech models. Once this process is complete, the resulting models are returned, via a communications channel 22, to speech recognition apparatus 14 to appropriate user sites 10_1, 10_2, ... 10_n. Communications channel 22 may be the same as communications channel 16, or may differ therefrom, as may be desired. The particular form of the respective communications channels forms no part of the present invention per se. The return, or redistribution, of retrained speech models may be effected at times when traffic on the communications channel is relatively light. Alternatively, the retrained speech models may be recorded on a CD-ROM or other portable storage device and delivered to a user site to replace the original speech models stored in speech recognition apparatus 14.

Thereafter, that is, after the original speech models that were stored in speech recognition apparatus 14 have been replaced by new or updated models, subsequent speech recognition of utterances from users of the subject class is carried out with a higher degree of confidence. Hence, updated HMM's (or other speech models) are used at the user's site to recognize the user's speech even though the user is of a class for which the original design of the speech recognition apparatus does not operate with a high degree of confidence.

Turning now to Figs. 2A-2B, there is illustrated a flow chart representing the overall operation of the system depicted in Fig. 1. The user operates

user input 12 to enter predetermined criteria information used to identify the class of which the user is a member. Step 32 represents examples of such criteria information, including the primary language spoken by the user, the user's gender, the user's age, the age of the user when he first learned the system target language, and the number of years the user has spoken the target language. In addition, speech samples consisting of group-specific registration sentences are spoken by the user and are included in the criteria information. As mentioned above, such registration sentences are predetermined words, phrases and sentences that the system has been programmed to recognize, that are designed to characterize the user's speech issues. For example, native Japanese speakers have a problem pronouncing the consonant 'L', so the registration sentences for this speaker group would include words containing 'L' to determine whether existing models will identify the user's utterance of this consonant. Hence, depending upon the dialect, accent and other features of the user's voice and speech pattern, speech recognition apparatus 14 is able to select best-matched models to represent these calibrated speech samples. As a result, rejection scores and correction data for this user are established, based upon the registration speech samples.

It will be appreciated that the criteria data represented in step 32 are intended to be representative examples and are not all-inclusive of the information used to identify the user's class.

In addition to entering criteria data, the user also enters utterances which are sampled and compared to a predetermined set of stored speech models by speech recognition apparatus 14, as represented by step 34 in Fig. 2. As mentioned previously, the speech recognition apparatus operates in a manner known to those of ordinary skill in the art, such as described in co-pending application (attorney's docket 450100-02964), to sense the user's utterances. The sensed utterance is sampled to extract therefrom identifiable speech features. These speech features are compared to the stored speech models and the optimal match between a sequence of features and the stored models is obtained. It is expected that the best-matched model nevertheless differs from the sampled feature sequence to the extent that an improved set of speech models is needed to optimize recognition performance. As represented by step 36, these models are downloaded from a suitable library, such as a read-only memory

device (e.g. a CD-ROM, a magnetic disk, a solid-state memory device, or the like) at the user's site.

Step 38 indicates that the speech recognition apparatus at the user's site monitors its performance to identify speech samples, extracted features and utterances that are recognized with a lower degree of confidence and, thus, require model-correction to improve the rejection score of those features. Such utterances, and their correction information, are stored; and as mentioned above, when traffic on the communications link between the user's site and the remote central site is light, these utterances and their correction data, together with the criteria data that identify the user's class are transferred, or uploaded to that central site. In one embodiment of the present invention, the samples that are uploaded to the central site are phonetically transcribed using a context-constrained (e.g. phono tactically-constrained) network at that site, as represented by step 40. That is, the system attempts to transcribe speech in a subword-by-subword manner and then link, or string those subwords together to form words or speech passages. As is known, context-constraints provide a broad set of rules that constrain allowable sequences of subwords (e.g. phonemes) to only those that occur in the target language. While step 40 is described herein as being carried out at the central site, it will be appreciated by those of ordinary skill in the art that this step may be performed at the user's site.

The transcribed samples are linked, or strung together using canonical (i.e. dictionary-supplied) subword spellings, resulting in words, as represented by step 42. An example of a canonical spelling is "sihks" or "sehks" to represent the numeral six. A fully shared network (i.e. connected so that all allowable spelling variants are considered) based upon both canonical spellings and/or subword transcriptions is built. Then, as depicted in step 44, the new data presented to this fully shared network is used to find transcriptions, or paths, through those models whose rejection scores are greater than the rejection scores for the initial transcriptions. This results in improved subword transcriptions that return higher rejection scores. Inquiry 46 then is made to determine if there are, in fact, a sufficient number of utterances of improved rejection scores. If so, step 56 is carried out to return the transcriptions created from such utterances as updated models for use in the lexica at those sites of users of the class from which the improved utterances were created. Since a

collection of speech samples and correction data are derived from the class of users, the number of such samples and correction data must be sufficiently large, that is, they must exceed a predetermined threshold number, before it is concluded that there are a sufficient number of improved utterances and speech models to be returned to the users.

If inquiry 46 is answered in the negative, the flowchart illustrated in Fig. 2 advances to step 48 which accumulates the number of underperforming utterances, that is, those utterances whose rejection scores are in the best-matched but low confidence range (i.e. those utterances in the gray area). When the accumulated number of such underperforming utterances exceeds a preset threshold, it is concluded that a sufficient number of underperforming utterances has been accumulated and either retraining of the set of speech which resulted in these Viterbi scores models or the derivation of a new class and corresponding speech models is initiated. As represented by step 50, these types of underperforming utterances are identified; and qualifying class member sites in the network shown in Fig. 1 are instructed to transfer, or upload to the central site the speech samples and correction data for those identified utterances. Consequently, the collection of such speech samples and correction data accumulates; and when a sufficient number are stored, the preset speech models are retrained, or retrained as a new class, as depicted in step 52, to result in a closer match to the utterances that had been identified as underperforming. Thereafter, step 54 redistributes the corrected set of speech models to those sites of users in the qualifying class. Hence, updated sets of corrected speech models are returned to the user sites for use in subsequent speech recognition.

Referring now to the flow chart shown in Figs. 3A-3B, there is illustrated a more detailed representation of the system operation in accordance with the present invention. Step 62, like step 32 discussed above in conjunction with Fig. 2A, is carried out by the user who enters criteria data by operating user input 12 at, for example, user site 10_1. Thus, as described above, the user enters criteria information, including the primary language spoken by the user, the user's gender, the user's age, the user's height , the users's weight, the age of the user when he first learned the system target language, the number of years the user has spoken the target language and speech samples consisting of registration sentences. Speech recognition

apparatus 14 operates by sequentially applying the user's speech samples, one at a time, to a library of stored speech models, as represented by step 64. The operation of the speech recognition apparatus advances to step 66 to find the best match between the given user's speech samples and the stored speech models.

Then, inquiry 68 is made to determine if this sample differs from the best-matched model by at least a predetermined amount. It is appreciated that if the user's speech sample is recognized with a high degree of confidence, rejection score of that speech sample is relatively high, this inquiry is answered in the negative; and the operation of the speech recognition apparatus returns to step 66. The loop formed of step 66 and inquiry 68 recycles with each user speech sample until the inquiry is answered in the affirmative, which occurs when the rejection score of the sample is low enough to be in the aforementioned gray area. At that time, the programmed operation of the speech recognition apparatus advances to step 70, whereat the speech sample which differs from the best-matched model by at least the predetermined amount, along with label information added to a new training corpus corresponding to this best model. At Step 71, an evolution is done to check whether the process of the speech sample is completed. If not, a next sample is retrieved. The identified speech sample is added to a new training set; and the programmed operation advances to step 72 which either creates a new speaker/.model class or trains and updates the existing class. The correction is stored for subsequent reuse and the new or trained models are distributed to the appropriate sites.

In the embodiment depicted in Fig. 3A, the operation of using phono tactics and other conventional speech recognition rules are used in speech recognition apparatus 14 to link, or string acoustic subwords together so as to recognize words, as opposed to this operation being carried out at the central site, described above in conjunction with Figs. 2A-2B. Step 74 depicts this speech recognition operation carried out at the user's site. For example, depending upon the user's class, as determined by his registration of criteria data, the word "six" will be recognized differently from a user whose dialect is from the northern part of United States than from a user whose dialect is from the South. The linking of acoustic subwords from a Northerner may appear as, e.g., "s"--"ih"--"k"--"s"; whereas the linking of acoustic subwords from a Southerner may appear as "s"--"eh"--"k"--"s". Depending upon the

registered class of the user, the utterances may or may not be recognized. Then, inquiry 76 (Fig. 3B) is made to determine if the recognized words are correct. For example, if the result of step 74 yields a rejection score within a range of relatively low confidence, the speech recognition apparatus will return to the user, such as by way of a visual or audio cue, the query: "do you mean...?" If the user replies in the negative, thus meaning that the words recognized by step 74 are not correct, inquiry 76 is answered in the negative and the operation of the speech recognition system returns to step 74. Similarly, if the word recognized by step 74 is displayed to the user, either visually or audibly, and the user corrects the displayed word, inquiry 76 is answered in the negative or the user gives up. The system cycles through the loop form of step 74 and inquiry 76 until the inquiry is answered in the affirmative.

Then, the acoustic subword strings (which form the recognized words) having the best rejection scores are collected in step 78 and transferred, or uploaded, in step 80 to the central database along with the identified speech samples, the corrected speech samples and the correction data used to obtain the corrected speech samples. Also transferred to the central database are the criteria data that had been entered by the user in step 62. Hence, identified speech samples, corrected speech samples, correction data and acoustic subword strings associated with the user's class, as identified by the user himself, are stored in the central database, as represented by step 82. This information is stored as a function of the user's class. For example, and consistent with the examples mentioned above, corrected speech samples, correction data and acoustic subword strings the system target language for French males in the age group of 30-40 years who have spoken the target language for 12-18 years are stored in one storage location; corrected speech samples, correction data and acoustic subword strings for Japanese females in the age group of 20-25 years who have spoken the target language for 5-10 years are stored in another location; and so on. Inquiry 84 then is made to determine if a sufficient number of underperforming utterances are stored in the database. That is, if the number of speech samples collected for French males in the 30-40 year age group who have spoken the system target language for 12-18 years exceeds a preset amount, inquiry 84 is answered in the affirmative. But, if this inquiry is entered in the negative, the programmed

operation performed at the central site returns to step 80; and the loop formed of steps 80 and 82 and inquiry 84 is recycled until this inquiry is answered in the affirmative.

Once a sufficient number of underperforming utterances have been stored in the database, step 86 is performed to retrain the set of standardized speech models that had been stored in central database 18 (Fig. 1) and that had been distributed initially to the speech recognition apparatus at the respective user sites. The retraining is carried out in a manner known to those of ordinary skill in the art, as mentioned above, and then step 88 is performed to download these new, updated speech models to those user sites at which the registered users (i.e. at which the particular class of users) are located. Preferably, the speech recognition apparatus at the user's site uses both the updated speech models and the original speech models stored thereat to determine respective best matches to new utterances. If subsequent utterances are determined to be better matched to the updated speech models, the original speech models are replaced by the updated models.

Therefore, by the present invention a set of stored speech models that might not function accurately to recognize speech of particular classes of users nevertheless may be updated easily and relatively inexpensively, preferably from a central source, to permit the successful operation of speech recognition systems. Speech samples, user profile data and feature data collected from, for example, a particular class of users are used, over a period of time, to update the speech models used for speech recognition. It is appreciated that the flow charts described herein represent the particular software and algorithms of programmed microprocessors to carry out the operation represented thereby. While the present invention has been particularly shown and described with reference to a preferred embodiment(s), it will be understood that various changes and modifications may be made without departing from the spirit and scope of this invention. It is intended that the appended claims be interpreted to cover the embodiments described herein and all equivalents thereto.